Where the Truth Lies: Explaining the Credibility of Emerging Claims on the Web and Social Media

Kashyap Popat Subhabrata Mukherjee Jannik Strötgen Gerhard Weikum
Max Planck Institute for Informatics
Saarland Informatics Campus, Saarbrücken, Germany
{kpopat,smukherjee,jstroetge,weikum}@mpi-inf.mpg.de

ABSTRACT

The web is a huge source of valuable information. However, in recent times, there is an increasing trend towards false claims in social media, other web-sources, and even in news. Thus, fact-checking websites have become increasingly popular to identify such misinformation based on manual analysis. Recent research proposed methods to assess the credibility of claims automatically. However, there are major limitations: most works assume claims to be in a structured form, and a few deal with textual claims but require that sources of evidence or counter-evidence are easily retrieved from the web. None of these works can cope with newly emerging claims, and no prior method can give user-interpretable explanations for its verdict on the claim's credibility.

This paper overcomes these limitations by automatically assessing the credibility of emerging claims, with sparse presence in web-sources, and generating suitable explanations from judiciously selected sources. To this end, we retrieve diverse articles about the claim, and model the mutual interaction between: the stance (i.e., support or refute) of the sources, the language style of the articles, the reliability of the sources, and the claim's temporal footprint on the web. Extensive experiments demonstrate the viability of our method and its superiority over prior works. We show that our methods work well for early detection of emerging claims, as well as for claims with limited presence on the web and social media.

Keywords

Credibility Analysis, Text Mining, Rumor and Hoax Detection

1. INTRODUCTION

Despite providing huge amounts of valuable information, the web is also a source of false claims in social media, other web-sources and even in news that quickly reach millions of users. Misinformation occurs in many forms: erroneous quoting of or reporting on politicians or companies, faked reviews about products or restaurants, made up news on celebrities, etc. Detecting false claims and validating credible ones is challenging, even for humans [11]. Moreover, beyond mere classification, explanations are crucial so that assessments can be interpreted.

WWW 2017 Companion, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4914-7/17/04.
http://dx.doi.org/10.1145/3041021.3055133

Claim: Solar panels drain the sun's energy, experts say
Assessment: False
Explanation: Solar panels do not suck up the Sun's rays of photons. Just like wind farms do not deplete our planet of wind. These renewable sources of energy are not finite like fossil fuels. Wind turbines and solar panels are not vacuums, nor do they divert this energy from other systems.

Table 1: A sample claim with assessment and explanation.

Within prior work on credibility analysis (e.g., [6, 16, 17, 18]), the important aspect of providing explanations for credibility assessments has not been addressed. In most works, the analysis focuses on structured statements and exhibits major limitations: (i) claims take the form of subject-predicate-object triples [24] (e.g., `Obama_BornIn_Kenya`), (ii) questionable values for the object are easy to identify [16, 17] (e.g., `Kenya`), (iii) conflicts and alternative values are easy to determine [42] (e.g., `Kenya` vs. `USA`) and/or (iv) domain-specific metadata is available (e.g., user metadata in online communities such as who-replied-to-whom) [11, 23].

In our own prior work [29], we addressed some of these limitations by assessing the credibility of *textual claims*: arbitrary statements made in natural language in arbitrary kinds of online communities or other web-sources. Based on automatically found evidence from the web, our method could assess the credibility of a claim. However, like all other prior works, we restricted ourselves to computing a binary verdict (true or false) without providing explanations. Moreover, we assumed that we could easily retrieve ample evidence or counter-evidence from a (static) snapshot of the web, disregarding the dynamics of how claims emerge, spread, and are supported or refuted (i.e., stance of a web-source towards the claim).

This paper overcomes the limitations of these prior works (including our own [29]). We assess the credibility of newly emerging and "long-tail" claims with sparse presence on the web by determining the *stance, reliability, and trend* of retrieved sources of evidence or counter-evidence, and by providing user interpretable *explanations* for the credibility verdict.

Table 1 shows an example for the input and output of our method. For the given example, our model assesses its credibility as *false*, and provides user-interpretable explanation in the form of informative snippets automatically extracted from an article published by a reliable web-source refuting this claim — exploiting the interplay between multiple factors to show the explanation.

Our method works as follows. Given a newly emerging claim in the form of a (long) sentence or a paragraph at time t, we first use a search engine to identify documents from diverse web-sources referring to the claim. We refer to these documents as *reporting articles*. For assessing the credibility of the emerging claim, our model captures the interplay between several factors: the *language* of the

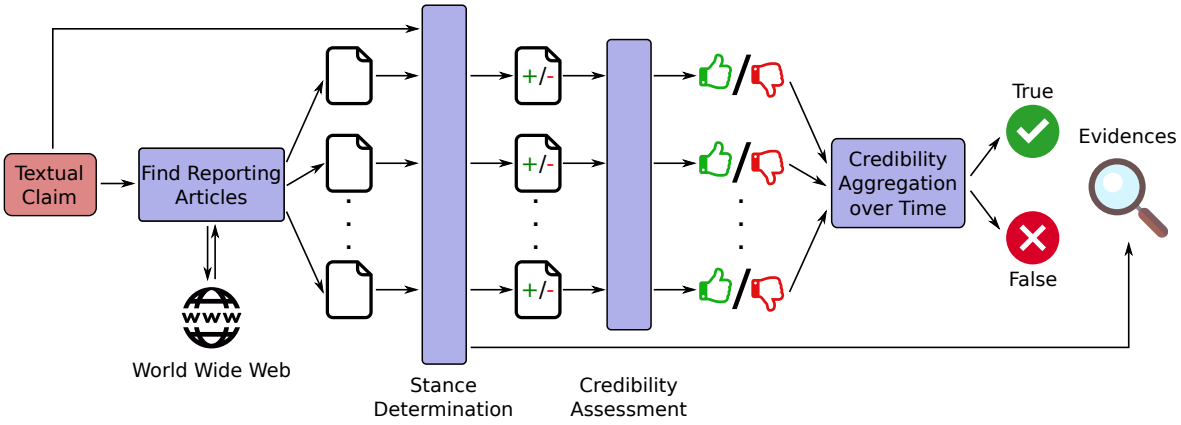

Figure 1: System framework for credibility assessment (+/- labels for articles indicate the stance i.e support/refute towards the claim).

reporting articles (e.g., bias, subjectivity, etc.), the *reliability* of the web-sources generating the articles, and the *stance* of the article towards the claim (i.e., whether it supports or refutes the claim). We propose two inference methods for the model: *Distant Supervision* and joint inference with a *Conditional Random Field* (CRF). The former approach learns all the factors sequentially, whereas the latter treats them jointly.

To tackle emerging claims and consider the temporal aspect, we harness the temporal footprint of the claim on the web, i.e., the dynamic trend in the timestamps of reporting articles that support or refute a claim. Finally, a joint method combines the content- and trend-aware models.

As evidence, our model extracts informative snippets from relevant reporting articles for the claim published by reliable sources, along with the stance (supporting or refuting) of the source towards the claim. Figure 1 gives a pictorial overview of the overall model. Extensive experiments with claims from the fact-checking website *snopes.com* and *wikipedia.com* demonstrate the strengths of our content-aware and trend-aware models by achieving significant improvements over various baselines. By combining them, we achieve the best performance for assessing the credibility of newly emerging claims. We show that our model can detect emerging false or true claims with a macro-averaged accuracy of 80% within 5 days of its origin on the web, with as low as 6 reporting articles per-claim.

Novel contributions of the paper can be summarized as:

- Exploring the interplay between factors like language, reliability, stance, and trend of sources of evidence and counter-evidence for credibility assessment of textual claims (*cf.* Section 3).
- Probabilistic models for joint inference over the above factors that give user-interpretable explanations (*cf.* Section 4).
- Experiments with real-world emerging and long-tail claims on the web and social media (*cf.* Section 5).

2. MODEL AND NOTATION

Our approaches based on distant supervision and CRF exploit the rich interaction taking place between various factors like source reliability and stance over time, article objectivity, and claim credibility for the assessment of claims. Figure 2 depicts this interaction. Consider a set of textual claims $\langle C \rangle$ in the form of sentences or short paragraphs, and a set of web-sources $\langle WS \rangle$ containing articles $\langle A^t \rangle$ that report on the claims at time t.

The following edges between the variables, and their labels, capture their interplay:

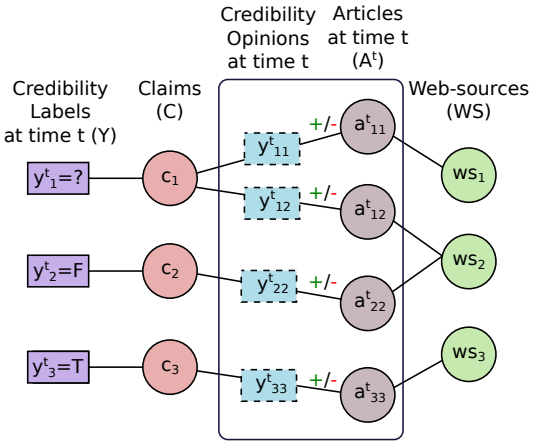

Figure 2: Factors for credibility analysis (+/- labels for edges indicate the article's stance i.e support/refute for the claim).

- Each claim $c_i \in C$ is connected to its reporting article $a_{ij}^t \in A^t$ published at time t.
- Each reporting article a_{ij}^t is connected to its web-source $ws_j \in WS$.
- For the joint CRF model, each claim c_i is also connected to the web-source ws_j that published an article a_{ij}^t on it at time t.
- Each article a_{ij}^t is associated with a random variable y_{ij}^t that depicts the *credibility opinion* (*True* or *False*) of the article a_{ij}^t (from ws_j) regarding c_i at time t — considering both the *stance* and *language* of the article.
- Each claim c_i is associated with a binary random variable y_i^t that depicts its *credibility label* at time t, where $y_i^t \in \{T, F\}$ (T stands for *True*, whereas F stands for *False*). y_i^t aggregates the individual credibility assessment y_{ij}^t of the articles a_{ij}^t for c_i at time t taking into account the reliability of their web-sources.

Problem statement: Given the labels of a subset of the claims (e.g., y_2^t for c_2, and y_3^t for c_3), our objective is to predict the credibility label of the newly emerging claim (e.g., y_1^t for c_1 at each time point t). The article set $\langle A^t \rangle$ and its predicted credibility label y^t for the newly emerging claim changes with time t as the evidence evolves.

3. CREDIBILITY ASSESSMENT FACTORS

We consider various factors for assessing the credibility of a textual claim. The following sections explain these factors.

Algorithm 1 Stance Determination Method

Input: Claim c_i and a corresponding reporting article a_{ij}^t at time t
Output: Stance scores (support & refute) of a_{ij}^t about c_i

1: Given a_{ij}^t, generate all possible snippets $\langle S \rangle$ of up to four consecutive sentences
2: Compute unigram & bigram overlap $\langle O \rangle$ of c_i with each snippet in $\langle S \rangle$
3: Remove snippets $\langle S^{'} \rangle$ with percentage overlap o_s with $c_i < \eta$
4: For each remaining snippet $s \in S \setminus S^{'}$, calculate its stance (support or refute) using a *stance classifier*
5: For each such snippet s, compute a combined score as the product of its stance probability and overlap score
6: Select top-k snippets $\langle S_{topK} \rangle$ based on the combined score
7: Return the average of stance support & refute scores of snippets in $\langle S_{topK} \rangle$

3.1 Linguistic Features

The credibility of textual claims heavily depends on the style in which it is reported. A true claim is assumed to be reported in an objective and unbiased language. On the other hand, highly subjective or sensationalized style of writing diminishes the credibility of a claim [24]. We use the same language features (F^L) (e.g., a set of assertive and factive verbs, hedges, report verbs, subjective and biased words etc.) as our prior work [29] to capture the linguistic style of the reporting articles:

- *Assertive and factive verbs* (e.g., “claim", “indicate") capture the degree of certainty to which a proposition holds.
- *Hedges* are the mitigating words (e.g., “may") which soften the degree of commitment to a proposition.
- *Implicative words* (e.g., “preclude") trigger presupposition in an utterance.
- *Report verbs* (e.g., “deny") emphasize the attitude towards the source of the information.
- *Discourse markers* (e.g., “could", “maybe") capture the degree of confidence, perspective, and certainty in the statements.
- Lastly, a lexicon of *subjectivity and bias* capture the attitude and emotions of the writer while writing an article.

3.2 Finding Stance and Evidence

In order to assess the credibility of a claim, it is important to understand whether the articles reporting the claim are supporting it or not. For example, an article from a reliable source like *truthorfiction.com* refuting the claim will make the claim less credible.

In order to understand the stance of an article, we divide the article into a set of snippets, and extract the snippets that are strongly related to the claim. This set of snippets helps in determining the overall score with which the article refutes or supports the claim. We compute both the support and refute scores, and use them as two separate features in our model.

The method for stance determination is outlined in Algorithm 1. Step 3 of the algorithm ensures that the snippets we consider are related to the claim. It removes snippets having overlap less than a threshold (η), where we consider all unigrams and bigrams for the overlap measure. In case all the snippets are removed in Step 3, we ignore the article. We varied η from 20% to 80% on withheld tuning data, and found $\eta = 40\%$ to give the optimal performance.

In Step 4, we use a *Stance Classifier* (described in the next section) to determine whether a snippet $s \in S \setminus S^{'}$ supports or refutes the claim. Let p_s^+ and p_s^- denote the corresponding support or refute probability of a snippet s coming from the classifier. We combine the stance probability of each snippet s with its overlap score o_s with the target claim: $\langle p_s^+ \times o_s, p_s^- \times o_s \rangle$. Then, we sort the snippets based on $\max(p_s^+ \times o_s, p_s^- \times o_s)$ and retrieve the top-k snippets S_{topK}. In our experiments (*cf.* Section 5), we set k to five. The idea is to capture the snippets which are highly related to the claim, and also have a strong refute or support probability.

Evidence: In the later stage, these snippets in $\langle S_{topK} \rangle$ are used as evidence supporting the result of our credibility classifier.

Feature vector construction: For each article a_{ij}^t, we average the two stance probabilities (for support and for refute) over the top-k snippets $s \in S_{topK}$ as two separate features:

$$F^{St}(a_{ij}^t) = \langle avg(\langle p_s^+ \rangle), avg(\langle p_s^- \rangle) \rangle.$$

3.2.1 Stance Classifier

Goal: Given a piece of text, the stance classifier should give the probability of how likely the text refutes or supports a claim based on the language stylistic features.

Data: Hoax debunking websites like *snopes.com*, *truthorfiction.com*, and *politifact.com* compile articles about contentious claims along with a manual analysis of the origin of the claim and its corresponding credibility label. We extract these analysis sections from such sources along with their manually assigned credibility labels (*true* or *false*). The *Stance Classifier* used in Step 4 of Algorithm 1 is trained using this dataset (withheld from the test cases later used in experiments). The articles confirming a claim are used as positive instances for the *“support"* class, whereas the articles debunking a claim are used as negative instances for the *“refute"* class.

Features: We consider all the unigrams and bigrams present in the training data as features, *ignoring all the named entities* (with part-of-speech tags NNP and NNPS). This is to prevent overfitting the model with popular entities (like “obama”, “trump”, “iphone”, etc.) which frequently appear in hoax articles.

Model: We use the L_2 regularized Logistic Regression (primal formulation) from the LibLinear package [8].

3.2.2 Training with Data Imbalance

Hoax debunking websites, by nature, mostly contain articles that *refute* rumors and urban legends. As a result, the training data for the stance classifier is imbalanced towards negative training instances from the “refute” class. For example, in *snopes.com*, this data imbalance is 2.8 to 1. In order to learn a balanced classifier, we adjust the classifier’s loss function by placing a large penalty[1] for mis-classifying instances from the positive or “support” class which boosts certain features from that class. The overall effect is that the classifier makes fewer mistakes for positive instances, leading to a more balanced classification.

3.3 Credibility-driven Source Reliability

Our prior work [29] used the PageRank and AlexaRank of web sources as a proxy for their reliability. However, these measures only capture the authority and popularity of the web-sources, and not their reliability from the credibility perspective. For instance, the satirical news website *The Onion* has a very high PageRank score (7 out of 10). Hence, we propose a new approach for measuring the source reliability that takes the authenticity of its articles into account. For each web-source, we determine the stance of its articles (regarding the respective claims) using the *Stance Classifier* explained above. A web-source is considered *reliable* if it contains articles that *refute false claims* and *support true claims*.

[1]We set the weight parameter in the LibLinear classifier to attribute a large penalty in the loss function for the class with less number of training instances.

Given a web-source ws_j with articles $\langle a_{ij}^t \rangle$ for claims $\langle c_i \rangle$ with corresponding credibility labels $\langle y_i^t \rangle$, we compute its reliability as:

$$reliability(ws_j) = \frac{\sum_{a_{ij}^t} \mathbf{1}\{St_{a_{ij}^t} = \text{'+'}, y_i^t = T\} + \sum_{a_{ij}^t} \mathbf{1}\{St_{a_{ij}^t} = \text{'-'}, y_i^t = F\}}{\text{cardinality}(\langle a_{ij}^t \rangle)}$$

where, $\mathbf{1}\{.\}$ is an indicator function which takes the value 1 if its argument is true, and 0 otherwise; $\{St_{a_{ij}^t} = \text{'+'}\}$ and $\{St_{a_{ij}^t} = \text{'-'}\}$ indicate that the article a_{ij}^t is supporting or refuting the claim, respectively. Thus, the first term in the numerator in the above equation counts the number of articles where a source *supports a true claim*, whereas the second term counts the number of articles where it *refutes a false claim*. Later, we use this reliability score of a source to weigh the credibility score of articles from a given source.

4. CREDIBILITY ASSESSMENT MODELS

We describe our different approaches for credibility assessment in the following sections.

4.1 Content-aware Assessment

Since the content-aware models are agnostic of time, we drop the superscripts t for all the variables in this section for notational brevity and better readability.

4.1.1 Model based on Distant Supervision

As credibility labels are available per-claim, and not per-reporting-article, our first approach extends the distant supervision based approach used in our prior work [29] by incorporating stance and improved source reliabilities. We attach the (observed) label y_i of each claim c_i to each article a_{ij} reporting the claim (i.e., setting labels $y_{ij} = y_i$). Using these $\langle y_{ij} \rangle$ as the corresponding training labels for $\langle a_{ij} \rangle$ with the corresponding feature vectors $\langle F^L(a_{ij}) \cup F^{St}(a_{ij}) \rangle$, we train an L_1-regularized logistic regression model on the training data along with the guard against data imbalance (*cf.* Section 3.2.2).

For any *test* claim c_i whose credibility label is unknown, and its corresponding reporting articles $\langle a_{ij} \rangle$, we use this *Credibility Classifier* to obtain the corresponding credibility labels $\langle y_{ij} \rangle$ of the articles. We determine the overall credibility label y_i of c_i by considering a weighted contribution of its *per-article* credibility probabilities, using the corresponding source reliability values as weights.

$$y_i = \arg\max_{l \in \{T,F\}} \sum_{a_{ij}} \big[reliability(ws_j) * Pr(y_{ij} = l) \big]$$

4.1.2 Joint Model based on CRF

The model described in the previous section learns the parameters for article stance, source reliability and claim credibility separately. A potentially more powerful approach is to capture the mutual interaction among these aspects in a probabilistic graphical model with joint inference, specifically a Conditional Random Field (CRF).

Consider all the web-sources $\langle WS \rangle$, articles $\langle A \rangle$, claims $\langle C \rangle$ and claim credibility labels $\langle Y \rangle$ to be nodes in a graph (*cf.* Figure 2). Let $\langle A_i \rangle$ be the set of all articles related to claim c_i. Each claim $c_i \in C$ is associated with a binary random variable $y_i \in Y$, where $y_i \in \{0, 1\}$ indicates whether the claim is *false* or *true*, respectively. We denote the *reliability* of web-source ws_j with α_j.

The CRF operates on the cliques of this graph. A clique, in our setting, is formed amongst a claim $c_i \in C$, a source $ws_j \in WS$ and an article $a_{ij} \in A$ about c_i found in ws_j. Different cliques are connected via the common sources and claims. There are as many cliques in the graph as the number of reporting articles. Let $\phi_{a_{ij}}(y_i, c_i, ws_j, a_{ij})$ be a potential function for the clique corresponding to a_{ij}. Each clique has a set of associated feature functions $F^{a_{ij}}$ with a weight vector θ. We denote the individual features and their weights as $f_k^{a_{ij}}$ and θ_k. The features are constituted by the stylistic, stance, and reliability features (*cf.* Sections 3.1, 3.2 & 3.3): $F^{a_{ij}} = \{\alpha_j\} \cup F^L(a_{ij}) \cup F^{St}(a_{ij})$.

We estimate the conditional distribution:

$$Pr(y_i | c_i, \langle ws_j \rangle, \langle a_{ij} \rangle; \theta) \propto \prod_{a_{ij}=1}^{|A_i|} \phi_{a_{ij}}(y_i, c_i, ws_j, a_{ij}; \theta)$$

The contribution of the potential of every clique $\phi_{a_{ij}}$ towards a claim c_i is weighed by the reliability of the source that takes its stance into account. Consider $\psi_{a_{ij}}(ws_j; \alpha_j, \theta_0)$ to be the potential for this reliability-stance factor. Therefore,

$$Pr(y_i | c_i, \langle ws_j \rangle, \langle a_{ij} \rangle; \theta) = \frac{1}{Z_i} \prod_{a_{ij}=1}^{|A_i|} \Big[\psi_{a_{ij}}(ws_j; \alpha_j, \theta_0) \times \phi_{a_{ij}}(y_i, c_i, ws_j, a_{ij}; \theta) \Big]$$

where,

$$Z_i = \sum_{y_i \in \{0,1\}} \prod_{a_{ij}=1}^{|A_i|} \Big[\psi_{a_{ij}}(ws_j; \alpha_j, \theta_0) \times \phi_{a_{ij}}(y_i, c_i, ws_j, a_{ij}; \theta) \Big]$$

is the normalization factor.

Assuming each factor takes the exponential family form, with features and weights made explicit:

$$\begin{aligned} &Pr(y_i | c_i, \langle ws_j \rangle, \langle a_{ij} \rangle; \theta) \\ &= \frac{1}{Z_i} \prod_{a_{ij}=1}^{|A_i|} \Big[\exp(\theta_0 \times \alpha_j) \times \exp(\sum_{k=1}^{K} \theta_k \times f_k^{a_{ij}}(y_i, c_i, ws_j, a_{ij})) \Big] \\ &= \frac{1}{Z_i} \exp(\theta_0 \times \sum_{a_{ij}=1}^{|A_i|} \alpha_j + \sum_{a_{ij}=1}^{|A_i|} \sum_{k=1}^{K} \theta_k \times f_k^{a_{ij}}(y_i, c_i, ws_j, a_{ij})) \\ &= \frac{1}{Z_i} \exp(\theta^T \cdot F^i) \end{aligned}$$

where, $F^i = [\sum_{a_{ij}=1}^{|A_i|} \alpha_j \sum_{a_{ij}=1}^{|A_i|} f_1^{a_{ij}} \sum_{a_{ij}=1}^{|A_i|} f_2^{a_{ij}} \cdots \sum_{a_{ij}=1}^{|A_i|} f_K^{a_{ij}}]$ and $\theta = [\theta_0\ \theta_1\ \theta_2\ \cdots \theta_K]$

We maximize the conditional log-likelihood of the data:

$$LL(\theta) = \sum_{i=1}^{|C|} \Big[\theta^T \cdot F^i \ - \ \log \sum_{y_i} \exp(\theta^T \cdot F^i) \Big] - \sigma ||\theta||_1$$

The L_1 regularization on the feature weights enforces the model to learn sparse features. The optimization for $\theta^* = \text{argmax}_\theta LL(\theta)$ is the same as that of logistic regression, with the *transformed* feature space. We use code from LibLinear [8] for optimization that implements trust region Newton method for large-scale logistic regression, with guard against data imbalance (*cf.* Section 3.2.2).

4.2 Trend-aware Assessment

Our hypothesis for this model is that the trend of articles supporting *true* claims increases much faster than the trend of refuting them over time; whereas, for *false* claims there is a trend of refuting them over time, rather than supporting them. To validate our hypothesis, we plot the cumulative number of supporting and refuting articles for each claim — aggregated over all the claims in our dataset — till each day $t \in [1 - 30]$ after the origin of a claim. As we can see from Figure 3, the cumulative support strength increases faster than the refute strength for true claims, and vice versa for false claims.

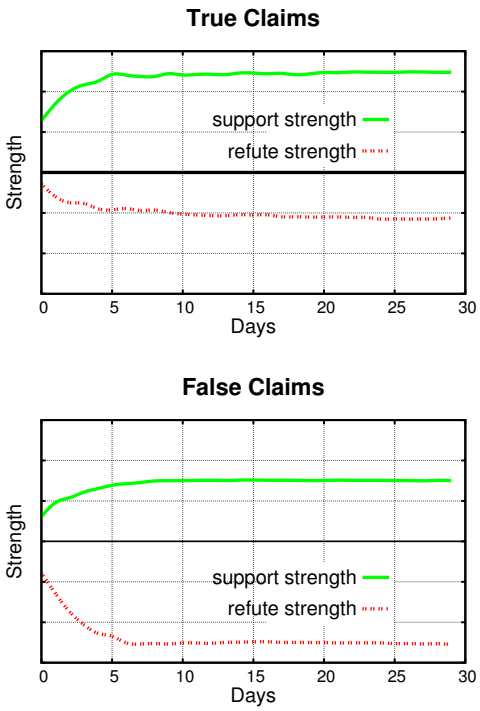

Figure 3: Trend of stance for *True* and *False* Claims.

We want to exploit this insight of evolving trends for credibility assessment of newly emerging claims. Thus, we revise our credibility assessment each day with new incoming evidence (i.e., articles discussing the claim) based on trend of support and refute.

In this approach, the credibility $Cr_{\text{trend}}(c_i, t)$ of a claim c_i at each day t is influenced by two components: (i) the *strength* of support and refute till time t (denoted by $q^+_{i,t}$ and $q^-_{i,t}$, respectively), and (ii) the *slope* of the trendline of support and refute (denoted by $r^+_{i,t}$ and $r^-_{i,t}$, respectively) till time t for the claim.

Let $\langle A^+_{i,t}\rangle$ and $\langle A^-_{i,t}\rangle$ denote the *cumulative* number of supporting and refuting articles for claim c_i till day t. The cumulative support and refute strength for the claim c_i till each day t is given by the mean of the stance scores, i.e., support and refute, denoted by p^+ and p^- (*cf.* Section 3.2), respectively — of all the articles reporting on the claim till that day, weighed by the reliability of their sources:

$$q^+_{i,t} = \frac{\sum_{a^t_{ij} \in A^+_{i,t}} p^+(a^t_{ij}) \times reliability(ws_j)}{|A^+_{i,t}|}$$

$$q^-_{i,t} = \frac{\sum_{a^t_{ij} \in A^-_{i,t}} p^-(a^t_{ij}) \times reliability(ws_j)}{|A^-_{i,t}|}$$

The slope of the trendline for the support and refute strength for the claim c_i till each day t is given by:

$$r^+_{i,t} = \frac{t \cdot \sum_{t'=1}^{t}(q^+_{i,t'} \cdot t') - \sum_{t'=1}^{t} q^+_{i,t'} \cdot \sum_{t'=1}^{t} t'}{t \cdot \sum_{t'=1}^{t} t'^2 - (\sum_{t'=1}^{t} t')^2}$$

$$r^-_{i,t} = \frac{t \cdot \sum_{t'=1}^{t}(q^-_{i,t'} \cdot t') - \sum_{t'=1}^{t} q^-_{i,t'} \cdot \sum_{t'=1}^{t} t'}{t \cdot \sum_{t'=1}^{t} t'^2 - (\sum_{t'=1}^{t} t')^2}$$

The trend based credibility score of claim c_i at time t aggregates the strength and slope of the trendline for support and refute as:

$$Cr_{\text{trend}}(c_i, t) = [q^+_{i,t} \cdot (1 + r^+_{i,t})] - [q^-_{i,t} \cdot (1 + r^-_{i,t})]$$

Total Claims	4856
True claims	1277 (26.3%)
False claims	3579 (73.7%)
Web articles	133272
Relevant articles	80421
Relevant web-sources	23260

Table 2: *Snopes* data statistics.

4.3 Content and Trend-aware Assessments

The content-aware approach analyzes the *language* of reporting articles from various sources. Whereas, the trend-aware approach captures the *temporal footprint* of the claim on the web for credibility assessment taking into account the trend of how various web-sources support or refute a claim over time. Hence, to take advantage of both the approaches, we combine their assessments for any claim c_i at time t as follows:

$$Cr_{comb}(c_i, t) = \alpha \cdot Cr_{\text{content}}(c_i, t) + (1 - \alpha) \cdot Cr_{\text{trend}}(c_i, t) \quad (1)$$

where, $Cr_{\text{content}}(c_i, t) = [Pr(y_i = \text{true})]$ (*cf.* Section 4.1) and $Cr_{\text{trend}}(c_i, t)$ are the credibility scores provided by the content-aware approach and trend-aware approach, respectively. $\alpha \in [0-1]$ denotes the combination weight.

5. EXPERIMENTS

5.1 Datasets

For assessing the performance of our approaches, we performed case studies on two real world datasets: (i) Snopes (*snopes.com*) and (ii) Wikipedia (*wikipedia.com*), which are made available online[2].

5.1.1 Snopes

Snopes is a well-known fact checking website that validates Internet rumors, e-mail forwards, hoaxes, urban legends, and other stories of unknown or questionable origin [38] receiving around 300,000 visits a day [28]. They typically collect these rumors and claims from *social media*, news websites, e-mails by users, etc. Each website article verifies a single claim, e.g., *"Clown masks have been banned in the United States, and wearing one can result in a $50,000 fine."*. The credibility of such claims are *manually* analyzed by Snopes' editors and labeled as *True* or *False*. For more details about the dataset, please refer to [29].

We collected these fact-checking articles from *Snopes* that are published until February 2016. For each claim c_i, we fired the *claim text* as a *query* to the Google search engine[1] and extracted the first three result pages (i.e., 30 articles) as a set of reporting articles $\langle a_{ij}\rangle$. We then crawled all these articles (using jsoup[3]) from their corresponding web-sources $\langle ws_j\rangle$. We removed search results from the *snopes.com* domain to avoid any kind of bias.

Statistics of the data crawled from *snopes.com* is given in Table 2. "Relevant" articles denote articles containing *at least* one snippet maintaining a stance (support or refute) about the target claim, as determined by our *Stance Classifier*. Similarly, relevant web-sources denote sources with at least one relevant article for any of the claims in our dataset.

[2] http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/impact/web-credibility-analysis/

[1] Our system has no dependency on Google. Other search engines or other means of evidence gathering could easily be used.

[3] https://jsoup.org/

	Hoaxes	Fictitious People
Total Claims	100	57
Web articles	2813	1552
Relevant articles	2092	1136
Relevant web-sources	1250	705

Table 3: Wikipedia data statistics.

Refute Class	Support Class
rumor, hoax, fake, false, satirical, fake news, spoof, fiction, circulate, not true, fictitious, not real, fabricate, reveal, can not, humor, misinformation, mock, unclear ...	review, editorial, accurate, speech, honor, display, marital, history, coverage, coverage story, read, now live, story, say, additional information, anticipate, examine ...

Table 4: Top contributing features for determining stance.

5.1.2 Wikipedia

Wikipedia contains a list of proven hoaxes[4] and fictitious people[5] (like fictional characters from novels). We used the same dataset as our prior work [29] of 100 hoaxes and 57 fictitious people. The ground-truth label for all of these claims is *False*. The statistics of the dataset is reported in Table 3. As described earlier, we used a search engine[1] to get a set of reporting articles for these claims by firing queries like "*<ENTITY> exists*" and "*<ENTITY> is genuine*". Similar to the previous case, we removed results from the *wikipedia.org* domain.

5.1.3 Time-series Dataset

As new claims emerge on the web, they are gradually picked up for reporting by various web-sources. To assess the performance of our trend-aware and combined approach for *emerging* claims, we require time-series data which mimics the behavior of emerging evidence (i.e., reporting articles) for newly emerged claims. Most of the prior works on rumor propagation dealt with online social networks (e.g., Twitter) [12, 45] where it is easy to trace the information diffusion. It is quite difficult to get such time-series data for the open web. In absence of any readily available dataset, we use a search engine to crawl the results.

Many of the Snopes articles contain the origin date of the claims. We were able to obtain 439 claims (54 *True* and 385 *False*) along with their date of origin on the web from Snopes. Now, to mimic the time-series behavior, we hit the Google search engine (using date restriction feature) and retrieved relevant reporting articles on a claim (first page of search results) on each day, starting from its day of origin to the next 30 days. We obtained 6000 *relevant* articles overall — as determined by our Stance Classifier. Using this time series dataset, the system's goal is to assess the credibility of a claim as soon as possible from its date of origin, given the set of reporting articles available in those initial days.

5.2 Stance and Source Reliability Assessment

To determine the stance of an article towards the claim, we trained our *Stance Classifier* (Section 3.2) using the *Snopes* data. The articles confirming (i.e., supporting) claims were taken as positive instances, whereas those debunking (i.e., refuting) claims were considered as negative instances. This trained model was used for determining the stance in both *Snopes* and *Wikipedia* datasets. We obtained **76.69%** accuracy with 10-fold cross-validation on labeled *Snopes* data for stance classification. Top contributing features for both classes are shown in Table 4.

As described in Section 3.3, we used the outcome of the stance determination algorithm to learn the reliability of various web-sources. The most reliable and most unreliable sources, as determined by our method, are given in Table 5.

Reliable	Non Reliable
wikipedia.org, *thatsfake.com*, *ibtimes.co.in*, *huffingtonpost.com*, *nydailynews.com*, *cnn.com*, *aljazeera.com* ...	*americannews.com*, *theonion.com*, *fox6now.com*, *huzlers.com*, *weeklyworldnews.com*, *dailycurrant.com* ...

Table 5: Top-ranked reliable and non-reliable sources.

5.3 Content-aware Assessment on Snopes

We perform 10-fold cross-validation on the claims by using 9-folds of the data for training, and the remaining fold for testing. The algorithm learned the *Credibility Classifier* and web-source reliabilities from the reporting articles and their corresponding sources present only in the training split. In case of a new web-source in test data, not encountered in the training data, its reliability score was set to 0.5 (i.e., equally probable of being reliable or not). We ignored all *Snopes*-specific references from the data and the search engine results in order to remove any training bias. For addressing the data imbalance issue (*cf.* Section 3.2.2), we set the penalty for the true class to 2.8 — given by the ratio of the number of *false* claims to *true* claims in the *Snopes* data.

5.3.1 Evaluation Measures

We report the overall accuracy of the model, Area-under-Curve (AUC) values of the ROC (Receiver Operating Characteristic) curve, precision, recall and F1 scores for the *False* claim class. *Snopes*, primarily being a hoax debunking website, is biased towards reporting *False* claims — the data imbalance being $2.8:1$. Hence, we also report the *per-class accuracy*, and the *macro-averaged accuracy* which is the average of *per-class* accuracy — giving equal weight to both classes irrespective of the data imbalance.

5.3.2 Baselines

We compare our approach with the following baselines implemented based on their respective proposed methods:

ZeroR[6]**:** A trivial baseline that always labels a claim as the class with the largest proportion in the dataset, i.e., *false* in our case.

Fact-finder Approaches: Approaches based on: (i) Generalized Sum [27], (ii) Average-Log [27], (iii) TruthFinder [42] and (iv) Generalized Investment [25] and (v) Pooled Investment [25]; implemented following the same method as suggested in [32].

Truth Assessment: Recent work on truth checking [24] utilizes the objectivity score of the reporting articles to find the truth. "Objectivity Detector" was constructed using the code[7] of [22]. A claim was labeled *true* if the sum of the objectivity scores of its reporting

[4] https://en.wikipedia.org/wiki/List_of_hoaxes#Proven_hoaxes

[5] https://en.wikipedia.org/wiki/List_of_fictitious_people

[6] https://weka.wikispaces.com/ZeroR

[7] Code and data available from: http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/impact/credibilityanalysis/

	Configuration	Overall Accuracy (%)	*True* Claims Accuracy (%)	*False* Claims Accuracy (%)	Macro-averaged Accuracy (%)	AUC	*False* Claims Precision	*False* Claims Recall	*False* Claims F1-Score
	CRF	**84.02**	71.26	**88.74**	80.00	0.86	0.89	**0.89**	**0.89**
Distant Supervision	**LG + ST + SR**	81.39	**83.21**	80.78	**82.00**	**0.88**	**0.93**	0.81	0.87
	ST + SR	79.43	80.12	79.22	79.67	0.86	0.92	0.79	0.85
	LG + ST	71.98	77.47	70.04	73.76	0.81	0.89	0.70	0.78
	Lang. + Auth.	71.96	75.43	70.77	73.10	0.80	0.89	0.71	0.79
	LG + SR	69.78	74.55	68.13	71.34	0.77	0.88	0.68	0.77
	ST	67.15	72.77	65.17	68.97	0.76	0.87	0.65	0.74
	LG	66.65	74.12	64.02	69.07	0.75	0.87	0.64	0.74

Table 6: Credibility classification results with different feature configurations (LG: language stylistic, ST: stance, SR: web-source reliability).

Configuration	Macro-averaged Accuracy (%)
ZeroR	50.00
Generalized Investment [25]	54.33
Truth Assessment [24]	56.06
TruthFinder [42]	56.91
Generalized Sum [27]	62.82
Pooled Investment [25]	63.09
Average-Log [27]	65.89
Lang. & Auth. [29]	73.10
Our Approach: CRF	**80.00**
Our Approach: Distant Supervision	**82.00**

Table 7: Performance comparison of our model vs. related baselines with 10-fold cross-validation.

articles was higher than the sum of the subjective scores, and *false* otherwise.

Our Prior Work (Lang. & Auth.): We also use our prior approach proposed in [29] which considers only the language of the reporting articles, and PageRank and AlexaRank based features for source authority to assess the credibility of claims.

5.3.3 Model Configurations

Along with the above baselines, we also report the results of our model with different feature configurations for linguistic style, stance, and credibility-driven web-source reliability:

- Models using only *language* (LG) features, only *stance* (ST) features, and their combination (LG + ST). These configurations use simple averaging of *per-article* credibility scores to determine the overall credibility of the target claim.
- The aggregation over articles is refined by considering the reliability of the web-source who published the article, considering *language and source reliability* (LG + SR), and *stance and source reliability* (ST + SR).
- Finally, all the aspects *language, stance* and *source reliability* (LG + ST + SR) are considered together.

5.3.4 Results

Table 7 shows the 10-fold cross-validation macro-averaged accuracy of our model against various baselines. As we can see from the table, our methods outperform all the baselines by a large margin. Table 6 shows the performance comparison of the different configurations. We can observe that using only language stylistic features (LG) is not sufficient; it is important to understand the stance (ST)

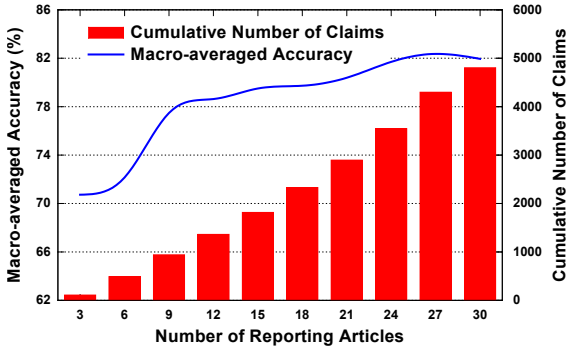

Figure 4: Performance on "long-tail" claims.

of the article as well. Considering stance along with the language boosts the *Macro-averaged Accuracy* by $\sim 5\%$ points.

The full model configuration, i.e., source reliability along with language style and stance features (LG + ST + SR), significantly boosts *Macro-averaged Accuracy* by $\sim 10\%$ points. High precision, recall and F1 scores for the *False* claim class show the strength of our model in detecting *False* claims. It also outperforms our prior work by a big margin which highlights the contribution of the *stance* and credibility-driven *source reliability* features.

We can observe from Table 6 that even though the overall accuracy of our CRF method is highest, it has comparatively a low performance on the true-claims class. Unlike the approach using Distant Supervision, the objective function in CRF is geared towards maximizing the *overall* accuracy, and therefore biased towards the false claims due to data imbalance. This persists even after adjusting the loss function during training to favor the positive class.

5.4 Handling "Long-tail" claims

In this experiment, we test the performance of our content-aware approach on "long-tail" claims that have only few reporting articles. We dissected the overall 10-fold cross-validation performance of our model based on the number of reporting articles of the claims. While calculating the performance, we considered *only* those claims which have $\leq k$ reporting articles, where $k \in \{3, 6, 9, \cdots 30\}$. Figure 4 shows the change in the *Macro-averaged Accuracy* for claims having different number of reporting articles. The Y-axis on the right hand side depicts the cumulative number of selected claims. The right-most bar in Figure 4 shows the performance of the LG + ST + SR configuration reported in Table 6. From the graph, we observe that our content-aware approach performs well even for "long-tail" claims having as few as 3 or 6 reporting articles.

Test Data	#Claims	Lang.+Auth. [29] Accuracy (%)	LG+ST+SR Accuracy (%)
WikiHoaxes	100	84	88
WikiFictitious People	57	66.07	82.14

Table 8: Accuracy of credibility classification on *Wikipedia*.

	Social Media	Web
Total claims	1566	1566
True claims	416	416
False claims	1150	1150
Relevant Web articles	6615	32668

Table 9: Data statistics: *Social Media* as source of evidence.

5.5 Content-aware Assessment on Wikipedia

To evaluate the generality of our content-aware approach, we train our model on the Snopes dataset, and test it on the *Wikipedia* dataset of hoaxes and fictitious people. The results in Table 8 demonstrate significant performance improvements over our prior work [29], and effectiveness of the *stance* and credibility-driven *source reliability* features in our model. Similar to the *Snopes* setting, we removed all references to Wikipedia from the data and search engine results. As we can see from the results, our system is able to detect hoaxes and fictitious people with high accuracy, although the claim descriptions here are stylistically quite different from those of *Snopes*.

5.6 Credibility Assessment of Emerging Claims

The goal of this experiment is to evaluate the performance of our approach with respect to the early assessment of newly emerging claims having sparse presence on the web. Using the time-series dataset (*cf.* Section 5.1.3), we assess the credibility of the emerging claims on each day t starting from their date of origin by considering the evidences (i.e., reporting articles) *only till day* t. We compare the macro-accuracy of the following approaches on each day t:

- *count-based approach:* In this approach, on each day t, we compare the cumulative number of supporting and refuting articles for a claim *till that day*. Stance is obtained using Algorithm 1 in Section 3.2. If the number of supporting articles is higher than the number of refuting ones, the claim is labeled *true*, and *false* otherwise.
- *trend-aware approach:* As described in Section 4.2, this analyzes the trend till day t to assess the credibility.
- *content-aware approach:* As described in Section 4.1, our model analyzes the content of relevant articles till day t and predicts the credibility of the claim.
- *content & trend-aware approach:* This combined approach considers credibility scores from both the models: content-aware and trend-aware (*cf.* Section 4.3). We varied the combination weight $\alpha \in [0-1]$ in steps of 0.1 on withheld development set, and found $\alpha = 0.4$ to give the optimal performance.

Results: Figure 5 shows the comparison of our approach with the baselines. As we observe in the figure, the count-based (baseline) approach performs the worst — thereby, ascertaining that simply counting the number of supporting / refuting articles is not enough. The best performance is achieved by the combined *content & trend-aware* approach. During the early days after a new claim has emerged, it leverages the trend to achieve the best performance. The results also highlight that we achieve *early detection* of emerging claims within $4-5$ days of its day of origin on the web with a high macro-averaged accuracy (ca. 80%). At the end of a month after the claim has emerged, all the approaches (except count-based) converge to similar results. The improvements in macro-accuracy for all of the respective approaches are statistically significant with p-value $< 2e-16$ using paired sample t-test.

Configuration	Overall Acc. (%)	*True* Claims Acc. (%)	*False* Claims Acc. (%)	Macro-averaged Acc. (%)
Social Media	76.12	77.34	75.66	76.50
Web	84.23	86.01	83.56	84.78

Table 10: Performance of credibility classification with different sources of evidence.

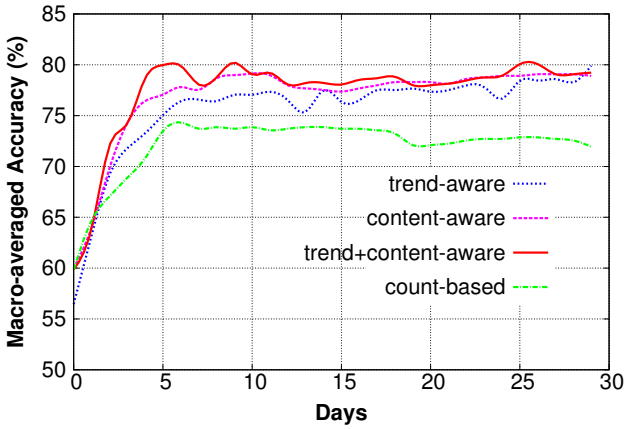

Figure 5: Comparison of macro-averaged accuracy for assessing the credibility of newly emerging claims.

5.7 Social Media as a Source of Evidence

Generally, social media is considered to be very noisy [1]. To test the reliability of social media in providing credibility verdicts for claims, we performed an additional experiment. We considered the following social media sites as potential sources of evidence: *Facebook*, *Twitter*, *Quora*, *Reddit*, *Wordpress*, *Blogspot*, *Tumblr*, *Pinterest*, *Wikia*. We selected the set of claims from the *Snopes* dataset (statistics are reported in Table 9) that had at least 3 reporting articles from the above mentioned sources. In the first configuration – *Social Media* – we used reporting articles only from these sources for credibility classification. In the second configuration – *Web* – we considered reporting articles from all sources on the web, including the social media sources. 10-fold cross-validation results for this task are reported in Table 10.

As we can observe from the results, relying *only* on social media results in a big drop of accuracy. Our system still performs decently. However, the system performance is greatly improved ($\sim 8\%$ points) by adding other sources of evidence from the web.

5.8 Evidence for Credibility Classification

Given a claim, our *Stance Classifier* extracts top-ranked snippets from the reporting articles along with their stance (*support* or *refute* probabilities). Combined with the verdict (*true* or *false*) from the *Credibility Classifier*, this yields evidence for the verdict. Table 11 shows examples of our model's output for some claims, along with the verdict and evidence. In contrast to all previous approaches, the assessment of our model can be easily interpreted by the user.

Claim	Verdict & Evidence
Titanium rings can be removed from swollen fingers only through amputation.	**[Verdict]:** False **[Evidence]:** A rumor regarding titanium rings maintains that ... This is completely untrue. In fact, you can use a variety of removal techniques to safely and effectively remove a titanium ring.
The use of solar panels drains the sun of energy.	**[Verdict]:** False **[Evidence]:** Solar panels do not suck up the Sun's rays of photons. Just like wind farms do not deplete our planet of wind. These renewable sources of energy are not finite like fossil fuels. Wind turbines and solar panels are not vacuums, nor do they divert this energy from other systems.
Facebook soon plans to charge monthly subscription fees to users of the social network.	**[Verdict]:**False **[Evidence]:** The rumor that Facebook will suddenly start charging users to access the site has become one of the social media era's perennial chain letters.
Soviet Premier Nikita Khrushchev was denied permission to visit Disneyland during a state visit to the U.S. in 1959.	**[Verdict]:** True **[Evidence]:** Soviet Premier Nikita Khrushchev's good-will tour of the United States in September 1959. While some may have heard of Khrushchev's failed attempt to visit Disneyland, many do not realize that this was just one of a hundred things that went wrong on this trip.
Between 1988 and 2006, a man lived at a Paris airport.	**[Verdict]:** True **[Evidence]:** Mehran Karimi Nasseri (born 1942) is an Iranian refugee who lived in the departure lounge of Terminal One in Charles de Gaulle Airport from 26 August 1988 until July 2006, when he was hospitalized for an unspecified ailment. His autobiography has been published as a book (The Terminal Man) and was the basis for the 2004 Tom Hanks movie The Terminal.

Table 11: Example claims with credibility verdict and automatically generated evidence from the Stance Classifier.

6. RELATED WORK

Our work is related to the following research areas:

Truth discovery: Truth discovery approaches [5, 6, 7, 9, 14, 15, 16, 18, 20, 25, 26, 42, 43, 44, 27] are mainly targeted towards resolving conflicts in multi-source data. These approaches assume that the input data has a structured representation and the conflicting values are already available. Work in [24] proposes a method to generate conflicting *values* or *fact candidates* from Web contents. They make use of linguistic features to detect the objectivity of the source reporting the fact. However, the work still depends on structured input in the form of Subject-Predicate-Object (SPO) triples, obtained by applying Open Information Extraction.

All the above approaches are limited to structured datasets with the main goal of conflict resolution amongst alternative fact candidates (or multi-source data). Our work addresses these limitations by proposing a general approach for credibility assessment for unstructured textual claims without requiring any alternative claims.

The method in [33] jointly estimates credibility of sources and correctness of the claims using the Probabilistic Soft Logic framework. However, unlike our approach, it does not consider the deeper semantic aspects of article language and the temporal footprint of claims on the web.

Credibility analysis within social media: [23] proposes a probabilistic graphical model jointly inferring user trustworthiness, language objectivity, and statement credibility. A similar approach in [22] identifies credible news articles, trustworthy news sources, and expert users. [41] works on extracting *Adverse Drug Reactions* from social media. Prior research for credibility assessment of social media posts exploits *community-specific* features for detecting rumors, fake, and deceptive content [4, 13, 30, 39, 40]. Temporal, structural, and linguistic features were used to detect rumors on Twitter in [12, 21]. [10] addresses the problem of detecting fake images in Twitter based on influence patterns and social reputation. A study on Wikipedia hoaxes is done in [11]. An algorithm for propagating trust scores in a network of claims, sources, and articles is proposed in [36].

All these approaches rely heavily on community-specific characteristics and are limited to online communities, network or social media. In contrast, we study credibility in an open domain setting without relying on such explicit signals.

Stance determination: Opinion mining methods for recognizing a speaker's stance in online debates are proposed in [34, 37]. Structural and linguistic features of users' posts are harnessed to infer their stance towards discussion topics in [35]. Temporal and textual information are exploited for stance classification over a sequence of tweets in [19]. In contrast to our work, these approaches are all tailored for debate forums.

Evidence detection: Approaches for Evidence Retrieval aim to find entire documents which can be used as evidence for a claim [2, 3]. In contrast, our model extracts informative textual snippets that support or refute a claim, instead of retrieving entire documents. The approach in [31] extracts the evidence from the document in the form of snippets. However, unlike our approach, it does not consider the stance of an article.

7. CONCLUSIONS

In this work, we propose approaches to leverage the stance, reliability and trend of sources of evidence and counter-evidence for credibility assessment of textual claims. Our experiments demonstrate that our system performs well on assessing the credibility of newly emerging claims within 4 to 5 days of its day of origin on the web with 80% accuracy; as well as for "long-tail" claims having as few as three reporting articles. Despite the fact that *social media* is very noisy, we show that our system can effectively harness evidence from such sources to validate or falsify a claim. In contrast to prior approaches, we provide explanations for our credibility verdict in the form of informative snippets from articles published by reliable sources that can be easily interpreted by the users. Experiments with data from the real-world fact-checking website *snopes.com* and reported cases of hoaxes and fictitious persons in Wikipedia demonstrate the superiority of our approaches over prior works.

8. ACKNOWLEDGMENTS

This research was partly supported by ERC Synergy Grant 610150 (imPACT).

9. REFERENCES

[1] T. Baldwin, P. Cook, M. Lui, A. MacKinlay, and L. Wang. How noisy social media text, how diffrnt social media sources? In *IJCNLP 2013*.

[2] P. Bellot, A. Doucet, et al. Report on inex 2013. *SIGIR Forum*, 47(2):21–32, Jan. 2013.

[3] M.-A. Cartright, H. A. Feild, and J. Allan. Evidence finding using a collection of books. In *BooksOnline 2011*.

[4] C. Castillo, M. Mendoza, and B. Poblete. Information credibility on twitter. In *WWW 2011*.

[5] X. L. Dong, L. Berti-Equille, and D. Srivastava. Integrating conflicting data: The role of source dependence. *Proc. VLDB Endow.*, 2(1):550–561, 2009.

[6] X. L. Dong, E. Gabrilovich, et al. Knowledge-based trust: Estimating the trustworthiness of web sources. *Proc. VLDB Endow.*, 8(9):938–949, 2015.

[7] X. L. Dong and D. Srivastava. Compact explanation of data fusion decisions. In *WWW 2013*.

[8] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *JMLR*, 2008.

[9] A. Galland, S. Abiteboul, A. Marian, and P. Senellart. Corroborating information from disagreeing views. In *WSDM 2010*.

[10] A. Gupta, H. Lamba, P. Kumaraguru, and A. Joshi. Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. In *WWW Companion 2013*.

[11] S. Kumar, R. West, and J. Leskovec. Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In *WWW 2016*.

[12] S. Kwon, M. Cha, K. Jung, W. Chen, and Y. Wang. Prominent features of rumor propagation in online social media. In *ICDM 2013*.

[13] T. Lavergne, T. Urvoy, and F. Yvon. Detecting fake content with relative entropy scoring. In *PAN 2008*.

[14] Q. Li, Y. Li, J. Gao, L. Su, B. Zhao, M. Demirbas, W. Fan, and J. Han. A confidence-aware approach for truth discovery on long-tail data. *Proc. VLDB Endow.*, 8(4), 2014.

[15] Q. Li, Y. Li, J. Gao, B. Zhao, W. Fan, and J. Han. Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation. In *SIGMOD 2014*.

[16] X. Li, X. L. Dong, K. Lyons, W. Meng, and D. Srivastava. Truth finding on the deep web: Is the problem solved? *PVLDB*, 6(2):97–108, 2012.

[17] X. Li, W. Meng, and C. Yu. T-verifier: Verifying truthfulness of fact statements. In *ICDE 2011*.

[18] Y. Li, Q. Li, J. Gao, L. Su, B. Zhao, W. Fan, and J. Han. On the discovery of evolving truth. In *KDD 2015*.

[19] M. Lukasik, P. K. Srijith, D. Vu, K. Bontcheva, A. Zubiaga, and T. Cohn. Hawkes processes for continuous time sequence classification: an application to rumour stance classification in twitter. In *ACL 2016*.

[20] F. Ma, Y. Li, Q. Li, M. Qiu, J. Gao, S. Zhi, L. Su, B. Zhao, H. Ji, and J. Han. Faitcrowd: Fine grained truth discovery for crowdsourced data aggregation. In *KDD 2015*.

[21] J. Ma, W. Gao, P. Mitra, S. Kwon, B. J. Jansen, K. Wong, and M. Cha. Detecting rumors from microblogs with recurrent neural networks. In *IJCAI 2016*.

[22] S. Mukherjee and G. Weikum. Leveraging joint interactions for credibility analysis in news communities. In *CIKM 2015*.

[23] S. Mukherjee, G. Weikum, and C. Danescu-Niculescu-Mizil. People on drugs: Credibility of user statements in health communities. In *KDD 2014*.

[24] N. Nakashole and T. M. Mitchell. Language-aware truth assessment of fact candidates. In *ACL 2014*.

[25] J. Pasternack and D. Roth. Knowing what to believe (when you already know something). In *COLING 2010*.

[26] J. Pasternack and D. Roth. Latent credibility analysis. In *WWW 2013*.

[27] J. Pasternack and D. Roth. Making better informed trust decisions with generalized fact-finding. In *IJCAI 2011*.

[28] D. Pogue. At snopes.com, rumors are held up to the light. `http://www.nytimes.com/2010/07/15/technology/personaltech/15pogue-email.html`, July 15, 2010. [Online; accessed 3-Oct-2016].

[29] K. Popat, S. Mukherjee, J. Strötgen, and G. Weikum. Credibility assessment of textual claims on the web. In *CIKM 2016*.

[30] V. Qazvinian, E. Rosengren, D. R. Radev, and Q. Mei. Rumor has it: Identifying misinformation in microblogs. In *EMNLP 2011*.

[31] R. Rinott, L. Dankin, et al. Show me your evidence - an automatic method for context dependent evidence detection. In *EMNLP 2015*.

[32] M. Samadi. *Facts and Reasons: Anytime Web Information Querying to Support Agents and Human Decision Making*. PhD thesis, Ph. D. Thesis, Computer Science Department, Carnegie Mellon University, Pittsburgh, PA, USA, 2015.

[33] M. Samadi, P. Talukdar, M. Veloso, and M. Blum. Claimeval: Integrated and flexible framework for claim evaluation using credibility of sources. In *AAAI 2016*.

[34] S. Somasundaran and J. Wiebe. Recognizing stances in online debates. In *ACL 2009*.

[35] D. Sridhar, L. Getoor, and M. Walker. Collective stance classification of posts in online debate forums. In *ACL Joint Workshop on Social Dynamics and Personal Attributes in Social Media 2014*.

[36] V. V. Vydiswaran, C. Zhai, and D. Roth. Content-driven trust propagation framework. In *KDD 2011*.

[37] M. A. Walker, P. Anand, R. Abbott, and R. Grant. Stance classification using dialogic properties of persuasion. In *NAACL HLT 2012*.

[38] Wikipedia. Snopes.com. `https://en.wikipedia.org/wiki/Snopes.com`. [Online; accessed 3-Oct-2016].

[39] Q. Xu and H. Zhao. Using deep linguistic features for finding deceptive opinion spam. In *COLING 2012*.

[40] F. Yang, Y. Liu, X. Yu, and M. Yang. Automatic detection of rumor on sina weibo. In *MDS 2012*.

[41] A. Yates, N. Goharian, and O. Frieder. Extracting adverse drug reactions from social media. In *AAAI 2015*.

[42] X. Yin, J. Han, and P. S. Yu. Truth discovery with multiple conflicting information providers on the web. *IEEE Trans. on Knowl. and Data Eng.*, 20(6):796–808, June 2008.

[43] B. Zhao, B. I. P. Rubinstein, J. Gemmell, and J. Han. A bayesian approach to discovering truth from conflicting sources for data integration. *Proc. VLDB Endow.*, 5(6):550–561, 2012.

[44] S. Zhi, B. Zhao, W. Tong, J. Gao, D. Yu, H. Ji, and J. Han. Modeling truth existence in truth discovery. In *KDD 2015*.

[45] A. Zubiaga, G. W. S. Hoi, M. Liakata, R. Procter, and P. Tolmie. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PLoS ONE*, 11(3), 2016.